

Simultaneous Clustering and Estimation of Heterogeneous Graphical Models

Will Wei Sun*, Botao Hao†, Yufeng Liu‡, Guang Cheng§

Abstract

We consider joint estimation of multiple graphical models arising from heterogeneous and high-dimensional observations. Unlike most previous approaches which assume that the cluster structure is given in advance, an appealing feature of our method is to learn cluster structure while estimating heterogeneous graphical models. This is achieved via a high dimensional version of Expectation Conditional Maximization (ECM) algorithm (Meng and Rubin, 1993). A joint graphical lasso penalty is imposed in the conditional maximization step to extract both homogeneity and heterogeneity components across all clusters. Our algorithm is computationally efficient due to fast sparse learning routines and can be implemented without unsupervised learning knowledge. The superior performance of our method is demonstrated by extensive experiments and its application to a Glioblastoma cancer dataset reveals some new insights in understanding the Glioblastoma cancer. In theory, a non-asymptotic error bound is established for the output directly from our high dimensional ECM algorithm, and it consists of two quantities: *statistical error* (statistical accuracy) and *optimization error* (computational complexity). Such a result gives a theoretical guideline in terminating our ECM iterations.

Key Words: Clustering, finite-sample analysis, high-dimensional ECM, heterogeneous graphical models, non-convex optimization.

*Assistant Professor, Department of Management Science, University of Miami, Coral Gables, FL 33146. Email: wsun@bus.miami.edu. Part of this research work was conducted when he was a Ph.D candidate at Purdue.

†Ph.D student, Department of Statistics, Purdue University, West Lafayette, IN 47906. E-mail: hao22@purdue.edu.

‡Professor, Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics, Carolina Center for Genome Sciences, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, NC 27599. E-mail: yfliu@email.unc.edu. Research Sponsored by NIH/NCI grant R01 CA-149569 and NSF grant DMS-1407241.

§Corresponding Author. Professor, Department of Statistics, Purdue University, West Lafayette, IN 47906. E-mail: chengg@purdue.edu. Research Sponsored by NSF CAREER Award DMS-1151692, DMS-1418042, and ONR N00014-15-1-2331.

1 Introduction

Graphical models have been widely employed to represent conditional dependence relationships among a set of variables. The structure recovery of an undirected Gaussian graph is known to be equivalent to recovering the support of its corresponding precision matrix (Lauritzen, 1996). In the situation where data dimension is comparable to or much larger than the sample size, the penalized likelihood method is proven to be an effective way to learn the structure of graphical models (Yuan and Lin, 2007; Friedman et al., 2008; Shojaie and Michailidis, 2010a,b). When observations come from several distinct subpopulations, a naive way is to estimate each graphical model separately. However, separate estimation ignores the information of common structure shared across different subpopulations, and thus can be inefficient in some real applications. For instance, in the glioblastoma multiforme (GBM) cancer dataset from The Cancer Genome Atlas Research Network (TCGA, 2008), Verhaak et al. (2010) showed that GBM cancer could be classified into four subtypes. Based on this cluster structure, it has been suggested that although the graphs across four subtypes differ in some edges, they share many common structures. In this case, the naive procedure can be suboptimal (Danaher et al., 2014; Lee and Liu, 2015). Such applications have motivated recent studies on joint estimation methods (Guo et al., 2011; Danaher et al., 2014; Lee and Liu, 2015; Qiu et al., 2016; Wang, 2015; Cai et al., 2016a; Peterson et al., 2015) that encourage common structure in estimating heterogeneous graphical models. However, all aforementioned approaches crucially rely on an assumption that the class label of each sample is known in advance.

For certain problems, prior knowledge of the class membership may be available. But this may not be the case for the massive data with complex and unknown population structures. For instance, in online advertising, an important task is to find the most suitable advertisement (ad) for a given user in a specific online context. This could increase the chance of users' favorable actions (e.g., click the ad, inquire about or purchase a product). In recent years, user clustering has gained increasing attention due to its superior performance of ad targeting. This is because users with similar attributes, such as gender, age, income, geographic information, and online behaviors, tend to behave similarly to the same ad (Yan et al., 2009). Moreover, it is very important to understand conditional dependence relationships among user attributes in order to improve ad targeting accuracy (Wang et al., 2015a). Such conditional dependence relationships are expected to share commonality across different groups (user homogeneity) while maintaining some levels of uniqueness within each group (user heterogeneity) (Jeziorski and Segal, 2015). In this online advertising application, previously mentioned joint estimation methods are no longer applicable as they need to know the user cluster structure in advance. Furthermore, with the data being continuously collected, the number of underlying user clusters grows with the sample size (Chen

et al., 2009). This provides another reason for simultaneously conducting user clustering and joint graphical model estimation, which is much needed in the era of big data.

Our contributions in this paper are two-fold. On the methodological side, we propose a general framework of **S**imultaneous **C**lustering **A**nd estimation **N** of heterogeneous graphical models (SCAN). SCAN is a likelihood based method which treats the underlying class label as a latent variable. Based on a high-dimensional version of Expectation Conditional Maximization (ECM) algorithm (Meng and Rubin, 1993), we are able to conduct clustering and sparse graphical model learning at the same time. In each iteration of the ECM algorithm, the expectation step performs cluster analysis by estimating missing labels and the conditional maximization step conducts feature selection and joint estimation of heterogeneous graphical models via a penalization procedure. With an iteratively updating process, the estimation for both cluster structure and sparse precision matrices becomes more and more refined. Our algorithm is computationally efficient by taking advantage of the fast sparse learning in the conditional maximization step. Moreover, it can be implemented in a user-friendly fashion, without the need of additional unsupervising learning knowledge.

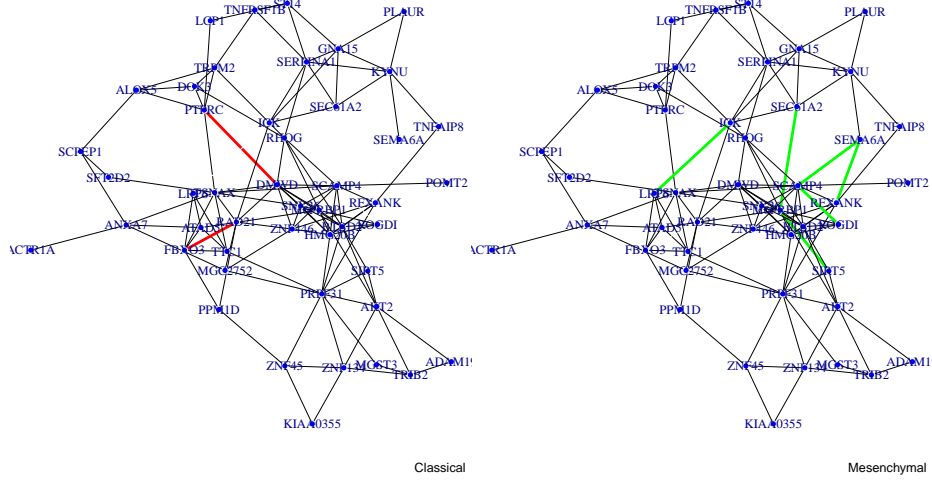
As a promising application, we apply the SCAN method on the GBM cancer dataset to simultaneously cluster the GBM patients and construct the gene regulatory network of each subtype. Our method greatly outperforms the competitors in clustering accuracy and delivers new insights in understanding the GBM disease. Figure 1 reports two gene networks estimated from the SCAN method. The black lines are links shared in both subtypes, and the thick red and thick green lines are uniquely presented in each subtype. Besides common edges of both subtypes, we have discovered some unique gene connections that were not found through separate estimation (Danaher et al., 2014; Lee and Liu, 2015). This new finding suggests further investigation on their possible impact on GBM disease. See Section 5 for more discussions.

On the theoretical side, we develop non-asymptotic statistical analysis for the output directly from the high dimensional ECM algorithm. This is nontrivial due to the non-convexity of the likelihood function. In this case, there is no guarantee that the sample-based estimator is close to the maximum likelihood estimator. Hence, we need to directly evaluate the estimation error in each iteration. Let Θ represent vectorized cluster means μ_k and precision matrices Ω_k , see (2.2) for a formal definition. Given an appropriate initialization $\Theta^{(0)}$, the finite sample error bound of the t -th step solution $\Theta^{(t)}$ consists of two parts:

$$\left\| \Theta^{(t)} - \Theta^* \right\|_2 \leq \underbrace{C \cdot \varepsilon(n, p, K, \Psi(\mathcal{M}))}_{\text{Statistical Error(SE)}} + \underbrace{\kappa^t \left\| \Theta^{(0)} - \Theta^* \right\|_2}_{\text{Optimization Error(OE)}}, \quad (1.1)$$

with high probability. Here, K is the number of clusters, $\Psi(\mathcal{M})$ measures the sparsity of cluster means and precision matrices, and $\kappa \in (0, 1)$ is a contraction coefficient. The above theoretical

Figure 1: Estimated gene networks corresponding to the Classical and Mesenchymal clusters from our SCAN method applying to the Glioblastoma Cancer Data. In each network, the black lines are the links shared in both groups. The red lines are the edges only appeared in the Classical cluster and the green lines are the edges only appeared in the Mesenchymal cluster.



analysis is applicable to any decomposable penalty used in the conditional maximization step.

The error bound (1.1) enables us to monitor the dynamics of estimation error in each iteration. Specifically, the optimization error decays geometrically with the iteration number t , while the statistical error remains the same when t grows. Therefore, the maximal number of iterations T is implied, beyond which the optimization error is dominated by the statistical error such that consequently the whole error bound is in the same order as the statistical error. In particular,

$$\sum_{k=1}^K \left(\left\| \mu_k^{(T)} - \mu_k^* \right\|_2 + \left\| \Omega_k^{(T)} - \Omega_k^* \right\|_F \right) = O_P \left(\underbrace{\sqrt{\frac{K^5 d \log p}{n}}}_{\text{Cluster means error}} + \underbrace{\sqrt{\frac{K^3 (Ks + p) \log p}{n}}}_{\text{Precision matrices error}} \right),$$

where d and s are the sparsity for a single cluster mean and precision matrix. This result indicates that, after T steps, the SCAN estimator will fall within statistical precision of the true parameter $\{\mu_k^*, \Omega_k^*\}$. It is worth mentioning that our theory allows the number of clusters K to diverge polynomially with the sample size, reflecting a typical big data scenario. When K is fixed and the cluster structure is known, our statistical rate for the precision matrix estimation under Frobenius norm achieves the optimal rate established in Cai et al. (2016b), i.e. $O_P(\sqrt{(s + p) \log p / n})$.

In the literature, a related line of research focuses on methodological developments of high-dimensional clustering. Pan and Shen (2007) and Sun et al. (2012) introduced regularized model-based clustering and regularized K -means clustering, and Zhou et al. (2009) proposed a network-based

clustering approach by imposing a graphical lasso to each individual precision matrix estimation. However, the regularized model-based clustering assumes an identical covariance matrix in each cluster, while the network-based clustering treats each graphical model estimation separately. As pointed out in [Danaher et al. \(2014\)](#) and [Lee and Liu \(2015\)](#), ignoring the network information of other clusters may lead to suboptimal graphical model estimation. During the submission of our paper, we are aware of an independent work by [Gao et al. \(2016\)](#) who also considered the multiple precision matrices estimation via a Gaussian mixture model. In particular, [Gao et al. \(2016\)](#) considered the fused lasso penalty on the precision matrices. However, no theoretical guarantee was provided in [Zhou et al. \(2009\)](#) and [Gao et al. \(2016\)](#). On the other hand, our SCAN method is more general than these existing methods since we allow the sparsity in both cluster mean and precision matrices. Most importantly, our theoretical analysis of the general SCAN framework sheds some lights on the behavior of these existing methods, see Remark 2.1 for more discussions. Moreover, in terms of the heterogeneous graphical model estimation, [Saegusa and Shojaie \(2016\)](#) proposed an interesting two-stage method which used hierarchical clustering to obtain cluster memberships and then estimated the multiple graphical models based on the attained cluster assignments. Clearly, the accuracy of their graphical model estimations heavily relies on the success of the hierarchical clustering in the first stage. However, a noticeable limitation of the hierarchical clustering is its sensitivity to the noise ([Narasimhan et al., 2006](#); [Balcan et al., 2014](#)), and it is unknown how the clustering error in the first stage would affect the estimation performance in the second stage. Our SCAN method unifies the clustering and graphical model estimation into a single ECM optimization and this enables us to quantify the estimation error in each iteration.

Another line of related work is the theoretical analysis of EM algorithm ([Balakrishnan et al., 2016](#); [Yi and Caramanis, 2015](#); [Wang et al., 2015b](#)). Specifically, [Balakrishnan et al. \(2016\)](#) studied the low-dimensional Gaussian mixture model, while [Wang et al. \(2015b\)](#) and [Yi and Caramanis \(2015\)](#) considered its high dimensional extension. However, their methods are not applicable for the estimation of heterogeneous graphical models due to the assumed identity covariance matrix. In fact, our consideration of the general covariance matrix demands more challenging technical analysis since simultaneous estimation of cluster means and covariance matrices induces a bi-convex optimization beyond the non-convexity of the EM algorithm itself. This also explains why ECM is needed instead of EM. To address these technical issues, key ingredients of our theoretical analysis are to bound the dual norm of the gradient of an auxiliary Q -function and employ nice properties of bi-convex optimization ([Boyd et al., 2011](#)) in the regularized M-estimation framework ([Negahban et al., 2012](#)). See Section 3 for more details.

In terms of notation, we use $[K]$ to denote the set $\{1, 2, \dots, K\}$. For a vector $\boldsymbol{\mu} \in \mathbb{R}^p$, $\|\boldsymbol{\mu}\|_2$ is its Euclidean norm. For a matrix $\mathbf{X} \in \mathbb{R}^{p_1 \times p_2}$, we denote $\|\mathbf{X}\|_F$ and $\|\mathbf{X}\|_2$ as its Frobenius norm and

spectral norm, respectively, and define its matrix max norm as $\|\mathbf{X}\|_{\max} = \max_{i,j} |X_{ij}|$ and its max induced norm as $\|\mathbf{X}\|_{\infty} = \max_{i=1,\dots,p_1} \sum_{j=1}^{p_2} |X_{ij}|$, which is simply the maximum absolute row sum of the matrix. For a square matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$, let $\sigma_{\min}(\mathbf{A})$ and $\sigma_{\max}(\mathbf{A})$ be its smallest and largest eigenvalue respectively and $|\mathbf{A}|$ be its determinant. For a sub-Gaussian random variable Z , we use $\|Z\|_{\psi_2}$ and $\|Z\|_{\psi_1}$ to denote its Orlicz norm. In detail, $\|Z\|_{\psi_2} = \sup_{p \geq 1} p^{-1/2} (\mathbb{E}|Z|^p)^{1/p}$, $\|Z\|_{\psi_1} = \sup_{p \geq 1} p^{-1} (\mathbb{E}|Z|^p)^{1/p}$. For two sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, $a_n \lesssim b_n$ refers to the case that $a_n \leq Cb_n$ for some uniform constant C . We write $\mathbf{1}(\cdot)$ as an indicator function. Throughout this paper, we use $C, C_1, C_2, \dots, D, D_1, D_2, \dots$ to denote generic absolute constants, whose values may vary at different places.

The rest of this article is organized as follows. Section 2 introduces heterogeneous graphical models and the SCAN method. Section 3 provides some statistical guarantees for the output directly from the SCAN method. Section 4 shows some simulation results, followed by a real data analysis on the Glioblastoma cancer data in Section 5. Section 6 gives some discussions. The appendix is devoted to the technical details of the main theorems and is available upon request.

2 Methodology

In this section, we introduce the SCAN method that simultaneously conducts high-dimensional clustering and estimation of heterogeneous graphical models.

2.1 Heterogeneous Graphical Models

We start our discussions from heterogeneous graphical models with known labels. Assume we are given K groups of data sets $\mathcal{A}_1, \dots, \mathcal{A}_K$ and the samples in the k -th group are generated i.i.d. from the following Gaussian distribution:

$$f_k(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (2\pi)^{-p/2} |\boldsymbol{\Sigma}_k|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right\}, k = 1, \dots, K. \quad (2.1)$$

Let $\boldsymbol{\Omega}_k = \boldsymbol{\Sigma}_k^{-1}$ be the k -th precision matrix with the ij -th entry ω_{kij} . For the k -th pair of parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k)$, i.e.,

$$\boldsymbol{\mu}_k = \begin{pmatrix} \mu_{k1} \\ \vdots \\ \mu_{kp} \end{pmatrix}, \boldsymbol{\Omega}_k = \begin{pmatrix} \omega_{k11} & \cdots & \omega_{k1p} \\ \vdots & \ddots & \vdots \\ \omega_{kp1} & \cdots & \omega_{kpp} \end{pmatrix},$$

we write $\boldsymbol{\Theta}_k := \text{vec}(\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k) = (\mu_{k1}, \dots, \mu_{kp}, \omega_{k11}, \dots, \omega_{kp1}, \dots, \omega_{k1p}, \dots, \omega_{kpp}) \in \mathbb{R}^{p^2+p}$ as its vectorized representation, and write the parameter of interest $\boldsymbol{\Theta}$ as

$$\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_K)^\top \in \mathbb{R}^{K(p^2+p)}. \quad (2.2)$$

In some cases, there may also exist some common structure across K precision matrices. [Danaher et al. \(2014\)](#) formulated the joint estimation of heterogeneous graphical models as

$$\operatorname{argmax}_{\boldsymbol{\Omega}_1, \dots, \boldsymbol{\Omega}_K \succ 0} \sum_{k=1}^K \sum_{\mathbf{x} \in \mathcal{A}_k} \log f_k(\mathbf{x}; \boldsymbol{\Theta}_k) - \mathcal{P}(\boldsymbol{\Omega}_1, \dots, \boldsymbol{\Omega}_K), \quad (2.3)$$

where $\mathcal{P}(\boldsymbol{\Omega}_1, \dots, \boldsymbol{\Omega}_K)$ is an entry-wise penalty which encourages both sparsity of each individual precision matrix and similarity among all precision matrices.

In practice, the cluster label is not always available. A probabilistic model is thus needed to accommodate the latent structure in the data. Assume the observation $\mathbf{x}_i; i = 1, \dots, n$, from unlabeled heterogeneous population has the underlying density

$$f(\mathbf{x}) = \sum_{k=1}^K \pi_k f_k(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where π_k is the probability that an observation \mathbf{x}_i belongs to the k -th subpopulation. Consider the penalized log-likelihood function for the *observed data*

$$\log \mathcal{L}(\boldsymbol{\Theta} | \mathbf{X}) := \frac{1}{n} \sum_{i=1}^n \log \left(\sum_{k=1}^K \pi_k f_k(\mathbf{x}_i; \boldsymbol{\mu}_k, (\boldsymbol{\Omega}_k)^{-1}) \right) - \mathcal{R}(\boldsymbol{\Theta}).$$

Our **S**imultaneous **C**lustering **A**nd estimation **N** (SCAN) method aims to solve

$$\max_{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Omega}_k} \log \mathcal{L}(\boldsymbol{\Theta} | \mathbf{X}). \quad (2.4)$$

For an illustration, we take

$$\mathcal{R}(\boldsymbol{\Theta}) = \underbrace{\lambda_1 \sum_{k=1}^K \sum_{j=1}^p |\mu_{kj}|}_{\mathcal{P}_1(\boldsymbol{\Theta})} + \underbrace{\lambda_2 \sum_{k=1}^K \sum_{i \neq j} |\omega_{kij}|}_{\mathcal{P}_2(\boldsymbol{\Theta})} + \underbrace{\lambda_3 \sum_{i \neq j} \left(\sum_{k=1}^K \omega_{kij}^2 \right)^{1/2}}_{\mathcal{P}_3(\boldsymbol{\Theta})}, \quad (2.5)$$

where $\mathcal{P}_1(\boldsymbol{\Theta})$ and $\mathcal{P}_2(\boldsymbol{\Theta})$ impose sparsity of the estimated cluster mean and precision matrix, and $\mathcal{P}_3(\boldsymbol{\Theta})$ encourages similarity among all estimated precision matrices. The above three tuning parameters can be tuned efficiently via BIC, see [Section 4.1](#).

Remark 2.1. It is worth mentioning that our SCAN method is applicable to penalty functions other than (2.5). For instance, the cluster mean penalty can be replaced by the group lasso penalty in [Sun et al. \(2012\)](#) or the ℓ_0 -norm penalty in [Shen et al. \(2012\)](#). The group graphical lasso penalty for the precision matrix estimation can be substituted by the structural pursuit penalty in [Zhu et al. \(2014\)](#) or the weighted bridge penalty in [Rothman and Forzani \(2014\)](#). As shown in [Section 2.2](#), only a slight modification of our algorithm is needed to accommodate other penalty functions.

We also note that SCAN reduces to the regularized model-based clustering (Pan and Shen, 2007) when $\lambda_2 = \lambda_3 = 0$, reduces to Zhou et al. (2009) when $\lambda_3 = 0$, and reduces to Gao et al. (2016) when $\lambda_1 = 0$. Consequently, the technical tools developed for the SCAN estimator in Section 3 are also applicable to these special cases.

2.2 ECM Algorithm

In this subsection, we introduce an efficient ECM algorithm to solve the general non-convex optimization problem in (2.4). The ECM replaces each M-step with an conditional maximization (CM) step in which each parameter $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Omega}_k$ is maximized separately, by fixing other parameters.

Denote the latent cluster assignment matrix as \mathbf{L} , where $L_{ik} = \mathbb{1}(\mathbf{x}_i \in \mathcal{A}_k)$; $i = 1, \dots, n$, $k = 1, \dots, K$. If the cluster label L_{ik} is available, the penalized log-likelihood function for the *complete data* can be formulated as

$$\log \mathcal{L}(\boldsymbol{\Theta} | \mathbf{X}, \mathbf{L}) := \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K L_{ik} \left[\log \pi_k + \log f_k(\mathbf{x}_i; \boldsymbol{\Theta}_k) \right] - \mathcal{R}(\boldsymbol{\Theta}).$$

In the expectation step, the conditional expectation of the penalized log-likelihood function is computed as

$$\mathbb{E}_{\mathbf{L} | \mathbf{X}, \boldsymbol{\Theta}^{(t-1)}} \left[\log \mathcal{L}(\boldsymbol{\Theta} | \mathbf{X}, \mathbf{L}) \right] = Q_n(\boldsymbol{\Theta} | \boldsymbol{\Theta}^{(t-1)}) - \mathcal{R}(\boldsymbol{\Theta}), \quad (2.6)$$

where $\mathcal{R}(\boldsymbol{\Theta})$ is the penalty in (2.5) and

$$Q_n(\boldsymbol{\Theta} | \boldsymbol{\Theta}^{(t-1)}) := \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K L_{\boldsymbol{\Theta}^{(t-1)}, k}(\mathbf{x}_i) \left[\log \pi_k + \log f_k(\mathbf{x}_i; \boldsymbol{\Theta}_k) \right], \quad (2.7)$$

with the class label being computed based on the parameter $\boldsymbol{\Theta}^{(t-1)}$ and $\pi_k^{(t-1)}$ obtained at the previous iteration, that is,

$$L_{\boldsymbol{\Theta}^{(t-1)}, k}(\mathbf{x}_i) = \frac{\pi_k^{(t-1)} f_k(\mathbf{x}_i; \boldsymbol{\Theta}_k^{(t-1)})}{\sum_{k=1}^K \pi_k^{(t-1)} f_k(\mathbf{x}_i; \boldsymbol{\Theta}_k^{(t-1)})}. \quad (2.8)$$

In the conditional maximization step, maximizing (2.6) with respect to $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Omega}_k$ yields the update of parameters. In particular, the update of π_k is given as

$$\pi_k^{(t)} = \sum_{i=1}^n \frac{L_{\boldsymbol{\Theta}^{(t-1)}, k}(\mathbf{x}_i)}{n}, \quad (2.9)$$

and the update of $\boldsymbol{\mu}_k$ is given in the following Lemma.

Lemma 2.2. Let $\boldsymbol{\mu}_k^{(t)} := \arg \max_{\boldsymbol{\mu}_k} Q_n(\boldsymbol{\Theta} | \boldsymbol{\Theta}^{(t-1)}) - \mathcal{R}(\boldsymbol{\Theta})$ and denote $n_k := \sum_{i=1}^n L_{\boldsymbol{\Theta}^{(t-1)},k}(\mathbf{x}_i)$. We have, for $j = 1, \dots, p$,

$$\mu_{kj}^{(t)} = \begin{cases} g_{1,j}(\mathbf{x}; \boldsymbol{\Theta}_k^{(t-1)}) - \frac{n\lambda_1}{n_k \omega_{kjj}^{(t-1)}} \text{sign}(\mu_{kj}^{(t-1)}) & \text{if } \left| \sum_{i=1}^n g_{2,j}(\mathbf{x}_i; \boldsymbol{\Theta}_k^{(t-1)}) \right| > \lambda_1; \\ 0 & \text{otherwise,} \end{cases}$$

where

$$g_{1,j}(\mathbf{x}; \boldsymbol{\Theta}_k^{(t-1)}) = \frac{\sum_{i=1}^n L_{\boldsymbol{\Theta}^{(t-1)},k}(\mathbf{x}_i) \left(\sum_{l=1}^p x_{il} \omega_{klj}^{(t-1)} \right)}{\omega_{kjj}^{(t-1)} n_k} - \frac{\sum_{l=1}^p \mu_{kl}^{(t-1)} \omega_{klj}^{(t-1)}}{\omega_{kjj}^{(t-1)}} + \mu_{kj}^{(t-1)},$$

$$g_{2,j}(\mathbf{x}_i; \boldsymbol{\Theta}_k^{(t-1)}) = L_{\boldsymbol{\Theta}^{(t-1)},k}(\mathbf{x}_i) \left(\sum_{l=1, l \neq j}^p (x_{il} - \mu_{kl}^{(t-1)}) \omega_{klj}^{(t-1)} + x_{ij} \omega_{kjj}^{(t-1)} \right).$$

Note that if the lasso penalty is replaced with other penalty functions, then the update formula of $\boldsymbol{\mu}_k^{(t)}$ in Lemma 2.2 can be modified accordingly. Given pseudo sample covariance matrix \tilde{S}_k , we are able to develop an update formula for $\boldsymbol{\Omega}_k$ by establishing its connection with joint estimation of heterogeneous graphical models (2.3).

Lemma 2.3. The solution of maximizing (2.6) with respect to $(\boldsymbol{\Omega}_1, \dots, \boldsymbol{\Omega}_K)$ is equivalent to

$$(\boldsymbol{\Omega}_1^{(t)}, \dots, \boldsymbol{\Omega}_K^{(t)}) := \arg \max_{\boldsymbol{\Omega}_1, \dots, \boldsymbol{\Omega}_K \succ 0} \sum_{k=1}^K n_k \left[\log \det(\boldsymbol{\Omega}_k) - \text{trace}(\tilde{S}_k \boldsymbol{\Omega}_k) \right] - \mathcal{R}(\boldsymbol{\Theta}), \quad (2.10)$$

where \tilde{S}_k is a pseudo sample covariance matrix defined as

$$\tilde{S}_k := \frac{\sum_{i=1}^n L_{\boldsymbol{\Theta}^{(t-1)},k}(\mathbf{x}_i) (\mathbf{x}_i - \boldsymbol{\mu}_k^{(t-1)})^\top (\mathbf{x}_i - \boldsymbol{\mu}_k^{(t-1)})}{\sum_{i=1}^n L_{\boldsymbol{\Theta}^{(t-1)},k}(\mathbf{x}_i)}.$$

The solution for (2.10) can be solved efficiently via the ADMM algorithm by slightly modifying the joint graphical lasso algorithm in Danaher et al. (2014). We summarize the high-dimensional ECM algorithm for solving the SCAN method in Table 1. Our algorithm is computationally efficient due to fast sparse learning routines shown in Lemma 2.2 and Lemma 2.3.

In all of our experiments, we obtain $(\boldsymbol{\mu}_k^{(0)}, \boldsymbol{\Omega}_k^{(0)})$ by random initialization, which is computationally efficient and practically reliable. In the theoretical study, we require the initialization to be of a constant distance to the truth. See Remark 3.9 for more discussions. Also, in the implementation, ECM step in Step 2 is terminated when the updated parameters are close to their previous values:

$$\sum_{k=1}^K \left\{ \frac{\|\boldsymbol{\mu}_k^{(t)} - \boldsymbol{\mu}_k^{(t-1)}\|_2}{\|\boldsymbol{\mu}_k^{(t)}\|_2} + \frac{\|\boldsymbol{\Omega}_k^{(t)} - \boldsymbol{\Omega}_k^{(t-1)}\|_2}{\|\boldsymbol{\Omega}_k^{(t)}\|_2} \right\} \leq 0.01.$$

Table 1: The SCAN Algorithm

Input: $\mathbf{x}_1, \dots, \mathbf{x}_n$, number of clusters K , tuning parameters $\lambda_1, \lambda_2, \lambda_3$.
Output: Cluster label \mathbf{L} , cluster mean $\boldsymbol{\mu}_k$ and precision matrix $\boldsymbol{\Omega}_k$.
Step 1: Initialize cluster mean $\boldsymbol{\mu}_k^{(0)}$, positive definite precision matrix $\boldsymbol{\Omega}_k^{(0)}$, and set $\pi_k^{(0)} = 1/K$, for each $k \in [K]$.
Step 2: Until some termination conditions are met, for iteration $t = 1, 2, \dots$
(a) E-step. Find the cluster assignment $L_{\boldsymbol{\Theta}^{(t-1)}, k}(\mathbf{x}_i)$ as in (2.8).
(b) CM-step. Given $L_{\boldsymbol{\Theta}^{(t-1)}, k}(\mathbf{x}_i)$, update $\pi_k^{(t)}$, $\boldsymbol{\mu}_k^{(t)}$, and $\boldsymbol{\Omega}_k^{(t)}$ in (2.9), Lemma 2.2, Lemma 2.3, respectively.

Remark 2.4. In the existing high-dimensional EM algorithms where covariance matrix is assumed to be an identity matrix (Wang et al., 2015b; Yi and Caramanis, 2015), sample-splitting procedures have been routinely used in the M-step in order to facilitate the theoretical analysis. Although it simplifies theoretical developments, such a sample-splitting procedure does not take advantage of full samples in the M-step and is hard to implement in practice. Our Algorithm 1 is able to avoid this sample-splitting step but still enjoys nice theoretical properties, see Corollary 3.13 for more discussions on its statistical guarantee.

3 Statistical Guarantee

In this section, we establish statistical guarantee for the SCAN estimator based on sample-based analysis of (2.7) and population-based analysis of (3.3). Here, we consider high-dimensional setting where $p \gg n$ and K is allowed to diverge with n .

We start by introducing some useful notation. Denote the index set of diagonal components of K precision matrices by

$$\mathcal{G} = \bigcup_{k=1}^K \mathcal{G}_k, \text{ with } \mathcal{G}_k = (k(p+1), k(2p+2), \dots, k(p^2+p)), \quad (3.1)$$

that is, $\boldsymbol{\Theta}_{\mathcal{G}} = (\omega_{111}, \dots, \omega_{1pp}, \dots, \omega_{K11}, \dots, \omega_{Kpp}) \in \mathbb{R}^{Kp}$. Let \mathcal{O} be the complete index set of $\boldsymbol{\Theta}$ and $\mathcal{G}^c = \mathcal{O} \setminus \mathcal{G}$ be the complement set of \mathcal{G} . Denote $\mathcal{U}_k := \{i : \mu_{ki}^* \neq 0\}$ where $\boldsymbol{\mu}_k^*$ is the true mean parameter, $\mathcal{V}_k := \{(i, j) : i \neq j, \omega_{kij}^* \neq 0\}$ where $\boldsymbol{\Omega}_k^*$ is the true precision matrix and $\mathcal{S}_1 = \bigcup_{k=1}^K \mathcal{U}_k$, $\mathcal{S}_2 = \bigcup_{k=1}^K \mathcal{V}_k$. Define $\Xi \subseteq \mathbb{R}^{K(p^2+p)}$ as some non-empty convex set of parameters. Denote the support space \mathcal{M} as

$$\begin{aligned} \mathcal{M} := \bigg\{ \mathbf{V} \in \Xi \mid & \mu_{ki} = 0 \text{ for all } i \notin \mathcal{S}_1, \\ & \omega_{kij} = 0 \text{ for all pairs } (i, j) \notin \mathcal{S}_2, k = 1 \dots, K \bigg\}, \end{aligned} \quad (3.2)$$

where \mathbf{V} follows the same definition manner of Θ defined in (2.2). Denote the following sparsity parameters: $s := \#\{(i, j) : \omega_{kij}^* \neq 0, i, j = 1 \dots p, i \neq j, k = 1, \dots, K\}$, $d := \#\{i : \mu_{ik}^* \neq 0, i = 1, \dots, p, k = 1, \dots, K\}$.

3.1 Population-Based Analysis

We define a corresponding population version of Q_n in (2.7) as

$$Q(\Theta' | \Theta) := \mathbb{E} \left[\sum_{k=1}^K L_{\Theta, k}(\mathbf{X}) [\log \pi'_k + \log f_k(\mathbf{X}; \Theta'_k)] \right]. \quad (3.3)$$

Without loss of generality, we assume the true prior probability $\pi_k^* = 1/K$ for each $k = 1, \dots, K$. Recall that the update of weight in (2.9) is independent of the updates of other parameters. Consequently, according to (2.1), maximizing $Q(\Theta' | \Theta)$ over (μ'_k, Ω'_k) is equivalent to maximizing

$$\sum_{k=1}^K \mathbb{E} \left[L_{\Theta, k}(\mathbf{X}) \left\{ \frac{1}{2} \log \det(\Omega'_k) - \frac{1}{2} (\mathbf{X} - \mu'_k)^\top \Omega'_k (\mathbf{X} - \mu'_k) \right\} \right]. \quad (3.4)$$

Clearly, the update of (μ'_l, Ω'_l) is independent of the update of (μ'_t, Ω'_t) for any $t \neq l$. This enables us to characterize the update of each pair of parameters separately. For any $k = 1, \dots, K$, define

$$M_{\mu'_k}(\Omega'_k) := \arg \max_{\mu'_k} Q(\Theta' | \Theta) \text{ and } M_{\Omega'_k}(\mu'_k) := \arg \max_{\Omega'_k} Q(\Theta' | \Theta).$$

We show in Lemma 3.1 that the population update of μ'_k is independent of Ω'_k , while the population update of Ω'_k is a function of μ'_k .

Lemma 3.1. For any $k = 1, \dots, K$, we have

$$M_{\mu'_k}(\Omega'_k) = [\mathbb{E}[L_{\Theta, k}(\mathbf{X})]]^{-1} \mathbb{E}[L_{\Theta, k}(\mathbf{X}) \mathbf{X}], \quad (3.5)$$

$$M_{\Omega'_k}(\mu'_k) = \mathbb{E}[L_{\Theta, k}(\mathbf{X})] \left[\mathbb{E}[L_{\Theta, k}(\mathbf{X}) (\mathbf{X} - \mu'_k)(\mathbf{X} - \mu'_k)^\top] \right]^{-1}. \quad (3.6)$$

The difficulty of simultaneous clustering and estimation can be characterized by the following *sufficiently separable condition*. Define $\mathcal{B}_\alpha(\Theta^*) := \{\Theta \in \Xi : \|\Theta - \Theta^*\|_2 \leq \alpha\}$.

Condition 3.2 (Sufficiently Separable Condition). Denote $W = \max_j W_j$, $W' = \max_j W'_j$, $W'' = \max_j W''_j$ with W_j, W'_j, W''_j defined in (S.4), (S.7) and (S.8), respectively. We assume K clusters are sufficiently separable such that given an appropriately small parameter $\gamma > 0$, it holds a.s.

$$L_{\Theta, k}(\mathbf{X}) \cdot L_{\Theta, j}(\mathbf{X}) \leq \frac{\gamma}{24(K-1) \sqrt{\max\{W, W', W''\}}}, \quad (3.7)$$

for each pair $\{(j, k), j, k \in [K], j \neq k\}$ and any $\Theta \in \mathcal{B}_\alpha(\Theta^*)$.

Condition 3.2 requires that K clusters are sufficiently separable in the sense that \mathbf{X} belongs to the k -th cluster with probability either close to zero or close to one such that $L_{\Theta,k}(\mathbf{X}) \cdot L_{\Theta,j}(\mathbf{X})$ is close to zero. In the special case that $K = 2$ and $\Omega_1^* = \Omega_2^* = \mathbf{1}_p$, Balakrishnan et al. (2016) requires $\|\mu_1^* - \mu_2^*\|_2$ is sufficiently large. Our Condition 3.2 extends it to general K and general precision matrices. Note that the condition (3.7) is related with the number of clusters K . As K grows, the clustering problem gets harder and hence a stronger sufficiently separable condition is needed.

The next lemma guarantees that the curvature of $Q(\cdot|\Theta)$ is similar to that of $Q(\cdot|\Theta^*)$ when Θ is close to Θ^* , which is a key ingredient in our population-based analysis.

Lemma 3.3 (*Gradient Stability*). Under Condition 3.2, the function $\{Q(\cdot|\Theta), \Theta \in \Xi\}$ satisfies,

$$\|\nabla Q(\Theta^*|\Theta) - \nabla Q(\Theta^*|\Theta^*)\|_2 \leq \tau \cdot \|\Theta - \Theta^*\|_2, \quad (3.8)$$

with parameter $\tau \leq \gamma/12$ for any $\Theta \in \mathcal{B}_\alpha(\Theta^*)$. The gradient $\nabla Q(\Theta^*|\Theta)$ is taken with respect to the first variable of $Q(\cdot|\cdot)$.

3.2 Sample-Based Analysis

In this section, we analyze the sample-base function Q_n , defined as the objective function in (2.7). We need one regularity condition to ensure that Q_n is strongly concave in a specific Euclidean ball.

Condition 3.4. There exist some positive constants β_1, β_2 such that $0 < \beta_1 < \min_{k \in [K]} \sigma_{\min}(\Omega_k^*) < \max_{k \in [K]} \sigma_{\max}(\Omega_k^*) < \beta_2$.

Lemma 3.5 verifies the restricted strong concavity condition of Q_n . Note that (3.9) corresponds to the restricted eigenvalue condition in sparse linear regression (Negahban et al., 2012).

Lemma 3.5 (*Restricted Strong Concavity*). Suppose that Condition 3.4 holds. Then for any $\Theta \in \mathcal{B}_\alpha(\Theta^*)$, with probability at least $1 - \delta$, each $\Theta' \in \mathbb{C} := \{\Theta' \mid \|\Theta' - \Theta^*\|_2 \leq 2\alpha\}$ satisfies

$$Q_n(\Theta'|\Theta) - Q_n(\Theta^*|\Theta) - \left\langle \nabla Q_n(\Theta^*|\Theta), \Theta' - \Theta^* \right\rangle \leq -\frac{\gamma}{2} \left\| \Theta' - \Theta^* \right\|_2^2, \quad (3.9)$$

with sufficiently large n , where $\gamma = c \cdot \min\{\beta_1, 0.5(\beta_2 + 2\alpha)^{-2}\}$ is the strong concavity parameter for some constant c .

Define $\mathcal{P}(\Theta) = M_1 \mathcal{P}_1(\Theta) + M_2 \mathcal{P}_2(\Theta) + M_3 \mathcal{P}_3(\Theta)$ for some positive constants M_1, M_2, M_3 . Let \mathcal{P}^* be the dual norm of \mathcal{P} , which is defined as $\mathcal{P}^*(\Theta) = \sup_{\mathcal{P}(\Theta') \leq 1} \langle \Theta', \Theta \rangle$. For simplicity, write $\|\cdot\|_{\mathcal{P}^*} = \mathcal{P}^*(\cdot)$.

Condition 3.6. For any fixed $\Theta \in \mathcal{B}_\alpha(\Theta^*)$, with probability at least $1 - \delta_1$,

$$\left\| \nabla Q_n(\Theta^*|\Theta) - \nabla Q(\Theta^*|\Theta) \right\|_{\mathcal{P}^*} \leq \varepsilon_1, \quad (3.10)$$

and with probability at least $1 - \delta_2$, we have

$$\left\| \left[\nabla Q_n(\Theta^* | \Theta) - \nabla Q(\Theta^* | \Theta) \right]_{\mathcal{G}} \right\|_2 \leq \varepsilon_2, \quad (3.11)$$

where \mathcal{G} is defined in (3.1). Here ε_1 and ε_2 are functions of $n, p, K, \delta_1, \delta_2$.

Intuitively, ε_1 and ε_2 quantify the difference between the population-based and sample-based conditional maximization step. Note that \mathcal{P} does not penalize diagonal elements of each precision matrix, thus

$$\left\| \nabla Q_n(\Theta^* | \Theta) - \nabla Q(\Theta^* | \Theta) \right\|_{\mathcal{P}^*} = \left\| \left[\nabla Q_n(\Theta^* | \Theta) - \nabla Q(\Theta^* | \Theta) \right]_{\mathcal{G}^c} \right\|_{\mathcal{P}^*}.$$

Therefore, we can split the statistical error by two parts: one coming from the sparse estimate of cluster means and non-diagonal terms in precision matrices (3.10), and another one coming from the estimate of diagonal terms of precision matrices (3.11). In Lemma S.1, ε_1 and ε_2 will be specifically calculated for our proposed SCAN penalty. In the high dimensional ECM algorithm, there is no explicit form for the CM-step update due to the existence of penalty term. This is a crucial difference from the low-dimensional EM algorithm in Balakrishnan et al. (2016). Fortunately, the decomposability of SCAN penalty enables us to quantify statistical errors by evaluating the gradient of Q -function.

3.3 Statistical Error versus Optimization Error

In this section, we provide the final theoretical guarantee for the high-dimensional ECM algorithm by combining the population and sample-based analysis.

Definition 3.7 (*Support Space Compatibility Constant*). For the support subspace $\mathcal{M} \subseteq \mathbb{R}^{K(p^2+p)}$ defined in (3.2), we define

$$\nu(\mathcal{M}) = \sup_{\Theta \in \mathcal{M} \setminus \{0\}} \frac{\mathcal{P}(\Theta)}{\|\Theta\|_2}. \quad (3.12)$$

As shown in Negahban et al. (2012), $\nu(\mathcal{M})$ quantifies the degree of compatibility between the penalty and the error norm over support space \mathcal{M} . For our SCAN penalty, $\nu(\mathcal{M})$ is specifically calculated in Lemma A.2.

We first provide a general theory that applies to any decomposable penalty, such as the group lasso penalty in Sun et al. (2012) and fused graphical lasso penalty in Danaher et al. (2014). The theoretical result of our SCAN penalty will be discussed in Corollary 3.13.

Theorem 3.8. Suppose Conditions 3.2, 3.4, 3.6 hold and Θ^* lies in the interior of Ξ . Let $\kappa = 6\tau/\gamma$, where τ, γ are calculated in Lemma 3.3 and Lemma 3.5. Consider our SCAN algorithm in Table 1

with initialization $\Theta^{(0)}$ falling into a ball $\mathcal{B}_\alpha(\Theta^*)$ for some constant radius $\alpha > 0$ and assume the tuning parameters satisfy $\lambda_1 = M_1\lambda_n^{(t)}$, $\lambda_2 = M_2\lambda_n^{(t)}$, $\lambda_3 = M_3\lambda_n^{(t)}$, and

$$\lambda_n^{(t)} = \varepsilon + \kappa^t \frac{\gamma}{\nu(\mathcal{M})} \left\| \Theta^{(t-1)} - \Theta^* \right\|_2. \quad (3.13)$$

If the sample size n is large enough such that $\varepsilon \leq (1 - \kappa)\gamma\alpha/(6\nu(\mathcal{M}))$, then $\Theta^{(t)}$ satisfies, with probability at least $1 - t\delta'$,

$$\left\| \Theta^{(t)} - \Theta^* \right\|_2 \leq \underbrace{\frac{6\nu(\mathcal{M})}{(1 - \kappa)\gamma} \varepsilon}_{\text{Statistical Error(SE)}} + \underbrace{\kappa^t \left\| \Theta^{(0)} - \Theta^* \right\|_2}_{\text{Optimization Error(OE)}}, \quad (3.14)$$

where $\delta' = \delta + \delta_1 + \delta_2$ with δ , δ_1 , δ_2 defined in Lemma 3.5 and Condition 3.6 and $\varepsilon = \varepsilon_1 + \varepsilon_2/\nu(\mathcal{M})$.

The above theoretical result suggests that the estimation error in each iteration consists *statistical error* and *optimization error*. From the definition of τ in Lemma 3.3, κ is less than 0.5 so that it is a contractive parameter. With a relatively good initialization, even though ECM algorithm may be trapped into a local optima after enough iterations, it can be guaranteed to be within a small neighborhood of the truth, in the sense of statistical accuracy. In addition, with a proper choice of δ' , the final probability $1 - t\delta'$ will converge to 1; see Corollary 3.13 for details.

Remark 3.9. To our limited knowledge, there is no existing literature to guarantee the global convergence of ECM algorithm in a general case. Compromisingly, we have to require some constraints on the initial value. In our framework, the only requirement for the initial value is to fall into a ball with constant radius to the truth. Such a condition has also been imposed in EM algorithms (Balakrishnan et al., 2016; Wang et al., 2015b; Yi and Caramanis, 2015) and can be fulfilled by some spectral-based initializations (Zhang et al., 2014).

Remark 3.10. In Theorem 3.8, we introduce an iterative turning procedure (3.13) which appeared in high dimensional regularized M -estimation (Negahban et al., 2012), and was also applied in Yi and Caramanis (2015) to facilitate their theoretical analysis.

The error bound in (3.14) measures the estimation error in each iteration. Here, optimization error decays geometrically with the iteration number t , while the statistical error remains the same when t grows. Therefore, this enables us to provide a meaningful choice of the maximal number of iterations T beyond which the optimization error is dominated by the statistical error such that the whole error bound is in the same order of the statistical error.

In the following corollary, taking the SCAN penalty as an example, we provide a closed form of the maximal number of iterations T and also an explicit form of the estimation error.

Condition 3.11. The largest element of cluster means and precision matrices are both bounded, that is, for some positive constants c_1 and c_2 ,

$$\|\boldsymbol{\mu}^*\|_\infty := \max_{k \in [K]} \|\boldsymbol{\mu}_k^*\|_\infty < c_1 \text{ and } \|\boldsymbol{\Omega}^*\|_{\max} := \max_{k \in [K]} \|\boldsymbol{\Omega}_k^*\|_{\max} < c_2.$$

Condition 3.12. Suppose that the number of clusters K satisfies $K^2 = o(p(\log n)^{-1})$.

Corollary 3.13. Suppose Conditions 3.2, 3.4, 3.11 and 3.12 hold. If sample size n is sufficiently large such that

$$n \geq \left(\frac{6(CK\|\boldsymbol{\Omega}^*\|_\infty + C'K^{1.5})(\sqrt{Kd} + \sqrt{Ks} + \sqrt{K}) + C''K^{1.5}\sqrt{p}}{(1-\kappa)\gamma\alpha} \right)^2 \log p,$$

and the iteration step t is large enough such that

$$t \geq T = \log_{1/\kappa} \frac{\|\boldsymbol{\Theta}^{(0)} - \boldsymbol{\Theta}^*\|_2}{\varphi(n, p, K)},$$

where $\varphi(n, p, K) = 6\tilde{C}((1-\kappa)\gamma)^{-1}\|\boldsymbol{\Omega}^*\|_\infty(\sqrt{Kd} + \sqrt{Ks + p})\sqrt{K^3 \log p/n}$ for some positive constant \tilde{C} , the optimization error in (3.14) is dominated by the statistical error, and

$$\sum_{k=1}^K \left(\|\boldsymbol{\mu}_k^{(T)} - \boldsymbol{\mu}_k^*\|_2 + \|\boldsymbol{\Omega}_k^{(T)} - \boldsymbol{\Omega}_k^*\|_F \right) \leq \frac{12\tilde{C}}{(1-\kappa)\gamma} \left(\underbrace{\|\boldsymbol{\Omega}^*\|_\infty \sqrt{\frac{K^5 d \log p}{n}}}_{\text{Cluster means error}} + \underbrace{\|\boldsymbol{\Omega}^*\|_\infty \sqrt{\frac{K^3 (Ks + p) \log p}{n}}}_{\text{Precision matrices error}} \right),$$

with probability converging to 1.

Remark 3.14. If K is fixed, the above upper bound reduces to

$$\sum_{k=1}^K \left(\|\boldsymbol{\mu}_k^{(T)} - \boldsymbol{\mu}_k^*\|_2 + \|\boldsymbol{\Omega}_k^{(T)} - \boldsymbol{\Omega}_k^*\|_F \right) \lesssim \left(\underbrace{\|\boldsymbol{\Omega}^*\|_\infty \sqrt{\frac{d \log p}{n}}}_{\text{Cluster means error}} + \underbrace{\|\boldsymbol{\Omega}^*\|_\infty \sqrt{\frac{(s + p) \log p}{n}}}_{\text{Precision matrices error}} \right). \quad (3.15)$$

Consider the class of precision matrix $\mathcal{Q} := \{\boldsymbol{\Omega} : \boldsymbol{\Omega} \succ 0, \|\boldsymbol{\Omega}\|_\infty \leq C_{\mathcal{Q}}\}$ as in Cai et al. (2016b). When $C_{\mathcal{Q}}$ does not depend on n, p , our rate $\sqrt{(s + p) \log p/n}$ in (3.15) is minimax optimal for estimating s -sparse precision matrices under Frobenius norm, see Remark 4 in Cai et al. (2016b). Moreover, our cluster mean error rate $\sqrt{d \log p/n}$ is minimax optimal for estimating d -sparse cluster means, see Wang et al. (2015b).

Remark 3.15. As a by-product, we establish the variable selection consistency of $\boldsymbol{\Omega}_k^{(T)}$, which ensures that our precision matrix estimator can asymptotically identify true connected links. Assume $\|\boldsymbol{\Omega}_k^*\|_\infty$ is bounded and the minimal signal in the true precision matrix satisfies $\omega_{\min} :=$

$\min_{(i,j) \in \mathcal{V}_k, k=1, \dots, K} w_{kij}^* > 2r_n$, where $r_n = (\sqrt{K^5 d} + \sqrt{K^3(Ks + p)})\sqrt{\log p/n}$. The latter condition is weaker than that assumed in [Guo et al. \(2011\)](#), where they require a constant lower bound of ω_{\min} . To ensure the model selection consistency, we threshold the precision matrix estimator $\mathbf{\Omega}_k^{(T)}$ such that $\tilde{\omega}_{kij} = \omega_{kij}^{(T)} \mathbb{1}\{|\omega_{kij}^{(T)}| > r_n\}$ as in [Bickel and Levina \(2008\)](#) and [Lee and Liu \(2015\)](#). See Theorem [S.2](#) in the online supplementary for some results on the selection consistency result.

4 Simulation Study

In this section, we discuss an efficient tuning parameter selection procedure and demonstrate the superior numerical performance of our method. We compare our algorithm with three clustering and graphical model estimation methods:

- Standard K -means clustering ([MacQueen, 1967](#)).
- Algorithm in [Zhou et al. \(2009\)](#) which applies graphical lasso for each precision matrix estimation.
- A two-stage approach which first uses K -means clustering to obtain the clusters and then applies joint graphical lasso ([Danaher et al., 2014](#)) to estimate precision matrices.

4.1 Selection of Tuning Parameters

In our simultaneous clustering and graph estimation formulation, four tuning parameters $\Lambda := \{K, \lambda_1, \lambda_2, \lambda_3\}$ need to be appropriately determined so that both the clustering and network estimation performance can be optimized. In our framework, the tuning parameters are selected through a BIC-type selection criterion.

For a set of tuning parameters $\Lambda := \{K, \lambda_1, \lambda_2, \lambda_3\}$, the BIC criterion is defined as

$$\text{BIC}(\Lambda) = -2\log \hat{L}(\Lambda) + \text{df}(\Lambda) \log(n), \quad (4.1)$$

where $\hat{L}(\Lambda)$ is the sample likelihood function and $\text{df}(\Lambda)$ is degrees of freedom of the model. Here, $\text{df}(\Lambda)$ can be approximated by the size of selected variables in the final estimator. Therefore, according to the Gaussian mixture model assumption, the BIC criterion in [\(4.1\)](#) can be computed as

$$-2 \sum_{i=1}^n \log \left(\sum_{k=1}^K \hat{\pi}_k f_k(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_k, (\hat{\boldsymbol{\Omega}}_k)^{-1}) \right) + \log(n) \sum_{k=1}^K \left\{ \|\hat{\boldsymbol{\mu}}_k\|_0 + \|\hat{\boldsymbol{\Omega}}_k\|_0 \right\},$$

where $\hat{\pi}_k, \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Omega}}_k$ are final updates from Algorithm [1](#). In our experiment, we choose the optimal set of parameters which minimize the BIC value in [\(4.1\)](#).

4.2 Illustration

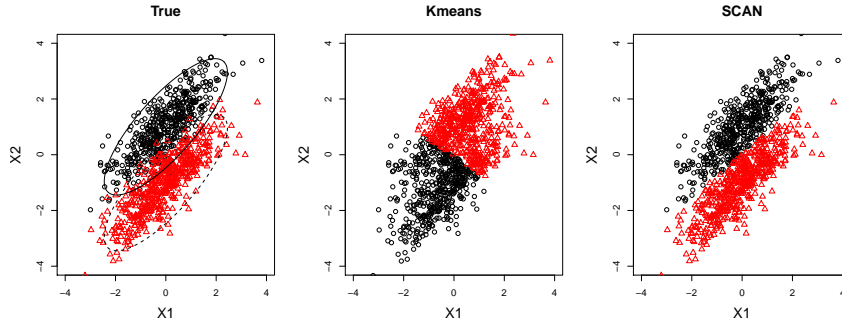
In this subsection, we demonstrate the importance of simultaneous clustering and estimation in improving both the clustering performance and the estimation accuracy of multiple precision matrices.

The simulated data consists of $n = 1000$ observations from 2 clusters, and among them 500 observations are from $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and the rest 500 observations are from $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu}_1 = (0, 1)^\top$, $\boldsymbol{\mu}_2 = (0, -1)^\top$, and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$

Since the covariance matrix $\boldsymbol{\Sigma}$ is not diagonal, the standard K -means algorithm which relies on the diagonal covariance assumption is expected to produce an unsatisfactory clustering result. This is illustrated in Figure 2 where K -means clustering clearly obtains wrong clusters. On the other hand, by incorporating the precision matrix estimation into clustering, our method is able to identify two correct clusters.

Figure 2: The first plot represents the true clusters shown in red and black in the example of Section 4.2. The middle and right plots show the clusters obtained from the standard K -means clustering (Kmeans) and our SCAN method.

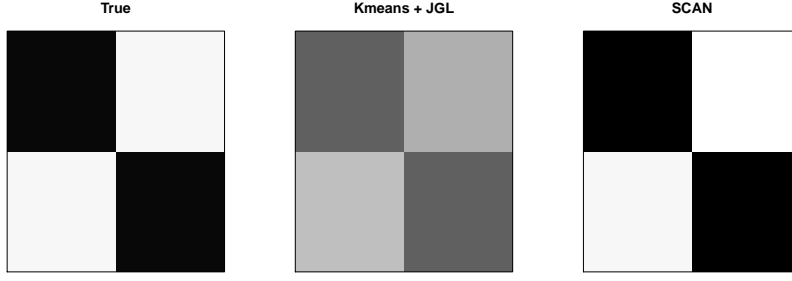


Moreover, Figure 3 illustrates the estimation performance of precision matrices based on the clusters estimated from the K -means clustering and our method. Clearly, our SCAN method delivers an estimator with improved accuracy when compared to the two stage method which applies joint graphical lasso (JGL) to the clusters obtained from the K -means clustering. This confirms that an accurate clustering is critical for the estimation performance of heterogeneous graphical models.

4.3 Simulations

In this subsection, we conduct extensive simulation studies to evaluate the performance of our algorithm. To assess the clustering performance of various methods, we compute the following

Figure 3: The true precision matrix and the estimated precision matrices from the two stage method (Kmeans + JGL) and our SCAN method in the example of Section 4.2.



clustering error (CE) which calculates the distance between an estimated clustering assignment $\hat{\psi}$ and the true assignment ψ of the sample data $\mathbf{X}_1, \dots, \mathbf{X}_n$ (Wang, 2010; Sun et al., 2012),

$$\text{CE}(\hat{\psi}, \psi) := \binom{n}{2}^{-1} \left| \{(i, j) : \mathbb{1}(\hat{\psi}(\mathbf{X}_i) = \hat{\psi}(\mathbf{X}_j)) \neq \mathbb{1}(\psi(\mathbf{X}_i) = \psi(\mathbf{X}_j)); i < j\} \right|,$$

where $|\mathcal{A}|$ is the cardinality of set \mathcal{A} . To measure the estimation quality, we calculate the precision matrix error (PME) and cluster mean error (CME)

$$\text{PME} := \frac{1}{K} \sum_{k=1}^K \left\| \hat{\boldsymbol{\Omega}}^{(k)} - \boldsymbol{\Omega}^{(k)} \right\|_2; \quad \text{CME} := \frac{1}{K} \sum_{k=1}^K \left\| \hat{\boldsymbol{\mu}}^{(k)} - \boldsymbol{\mu}^{(k)} \right\|_2.$$

Finally, to compare the variable selection performance, we compute the true positive rate (TPR, percentage of true edges selected) and the false positive rate (FPR, percentage of false edges selected)

$$\begin{aligned} \text{TPR} &:= \frac{1}{K} \sum_{k=1}^K \frac{\sum_{i < j} \mathbb{1}(\omega_{kij} \neq 0, \hat{\omega}_{kij} \neq 0)}{\sum_{i < j} \mathbb{1}(\omega_{kij} \neq 0)}, \\ \text{FPR} &:= \frac{1}{K} \sum_{k=1}^K \frac{\sum_{i < j} \mathbb{1}(\omega_{kij} = 0, \hat{\omega}_{kij} \neq 0)}{\sum_{i < j} \mathbb{1}(\omega_{kij} = 0)}. \end{aligned}$$

In the simulation, we considered a three-class problem and generated a 5-block tridiagonal precision matrix with 100 features for the precision matrix. To allow the similarity of precision matrices, we set the off-diagonal entry of $\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2, \boldsymbol{\Omega}_3$ as $\eta, 0.9\eta$, and 1.1η , respectively. The diagonal entries of $\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2$, and $\boldsymbol{\Omega}_3$ were all 1. The 300 samples were generated as follows. First, the cluster membership Y_i 's were uniformly sampled from $\{1, 2, 3\}$. Given the cluster label, we generated each sample $\mathbf{X}_i \sim \mathcal{N}(\boldsymbol{\mu}(Y_i), \boldsymbol{\Omega}(Y_i))$. Here, the cluster mean $\boldsymbol{\mu}(Y_i)$ is sparse, where its first 10 variables are of the form $(\mu \mathbf{1}_5^\top, -\mu \mathbf{1}_5^\top)^\top \mathbb{1}(Y_i = 1) + \mu \mathbf{1}_{10} \mathbb{1}(Y_i = 2) + (-\mu \mathbf{1}_5^\top, -\mu \mathbf{1}_5^\top)^\top \mathbb{1}(Y_i = 3)$, with $\mathbf{1}_5$ being a 5-dimensional vector of all ones, and its last $p - 10$ variables are zeros. We considered 3 simulation models with varying choices of μ and η :

- Model 1: $\mu = 0.8$ and $\eta = 0.3$,
- Model 2: $\mu = 1$ and $\eta = 0.3$,
- Model 3: $\mu = 1$ and $\eta = 0.4$.

Here, μ controls the separability of three clusters with larger μ corresponding to an easier clustering problem, and η represents the similarity level of precision matrices across clusters.

In the experiment, our method selected the tuning parameters via the BIC criterion in Section 4.1. For a fair comparison, we also used the same tuning parameters λ_1, λ_2 in Zhou et al. (2009), and the same λ_2, λ_3 in the joint graphical lasso penalty of the two-stage approach. We repeated the procedure 10 times and reported the averaged clustering errors, estimation errors, and variable selection errors for each method in Table 2. As shown in Table 2, the standard K -means clustering method has the largest clustering error due to a violation of its diagonal covariance matrix assumption. As the off-diagonal entries of the precision matrices increase from Model 2 to Model 3, the clustering error of standard K -means clustering increases dramatically, which leads to an increased estimation error of the precision matrices in the K -means+JGL method. In Model 1 and Model 3, the method of Zhou et al. (2009) improves the clustering performance of the standard K -means by using a graphical lasso in the precision matrix estimation. However, it obtains a relatively large precision matrix estimation error since it ignores the similarity across different precision matrices. In contrast, our SCAN algorithm achieves the best clustering accuracy and best precision matrix estimation accuracy for all the three simulation models. This is due to our simultaneous clustering and estimation strategy as well as the consideration of similarity of precision matrices across clusters. This experiment shows that a satisfactory clustering algorithm is critical to achieve accurate estimations of heterogeneous graphical models, and alternatively good estimation of the graphical model can also improve the clustering performance. This explains the success of our simultaneous method in terms of both clustering and graphical model estimation.

5 Glioblastoma Cancer Data Analysis

In this section, we apply our simultaneous clustering and graphical model estimation method to a Glioblastoma cancer dataset. We aim to cluster the glioblastoma multiforme (GBM) patients and construct the gene regulatory network of each subtype in order to improve our understanding of the GBM disease.

The raw gene expression dataset measures 17814 levels of mRNA expression of 482 GBM patients. Each patient belongs to one of four subgroups of GBM: Classical, Mesenchymal, Neural, and Proneural (Verhaak et al., 2010). In our analysis, we selected two subtypes, Classical (127

Table 2: The clustering errors (CE), cluster mean errors (CME), precision matrix errors (PME), true positive rates (TPR1) and false positive rates (FPR1) of cluster mean estimation, true positive rates (TPR2) and false positive rates (FPR2) of precision matrix estimation of four methods in the simulations of Section 4.3. The minimal clustering error and minimal estimation error in each simulation are shown in bold.

Models	Methods	CE	CME	PME	TPR1/FPR1	TPR2 /FPR2
Model 1 $\mu = 0.8$ $\eta = 0.3$	<i>K</i> -means	0.112	1.79	NA	1 /1	NA / NA
	Zhou et al. (2009)	0.104	1.64	10.93	0.97 /0.16	0.96 /0.1
	<i>K</i> -means + JGL	0.112	1.79	8.05	1 /1	0.99 /0.007
	SCAN	0.039	1.30	7.52	1 /0.15	1 /0.006
Model 2 $\mu = 1$ $\eta = 0.3$	<i>K</i> -means	0.015	1.30	NA	1 /1	NA / NA
	Zhou et al. (2009)	0.024	1.38	10.42	0.99 /0	0.97 /0.099
	<i>K</i> -means + JGL	0.015	1.30	7.55	1 /1	0.999 /0.006
	SCAN	0.007	1.29	7.50	1 /0	0.999 /0.006
Model 3 $\mu = 1$ $\eta = 0.4$	<i>K</i> -means	0.226	3.68	NA	1/1	NA / NA
	Zhou et al. (2009)	0.134	2.33	11.62	0.99 /0.24	0.996 /0.154
	<i>K</i> -means + JGL	0.226	3.68	10.53	1 /1	0.991 /0.04
	SCAN	0.070	1.97	8.57	0.97 /0.23	0.998 /0.04

samples) and Mesenchymal (145 samples), that had the largest sample sizes. Although they are biologically different, these two subtypes share many similarities since they are both GBM diseases. For our analysis, we considered the 840 signature genes established by Verhaak et al. (2010). Following the preprocess procedures in Lee and Liu (2015), we excluded the genes with no subtype information or the genes with missing values. We then applied the sure independence screening analysis (Fan and Lv, 2008) to finally include 50 genes in our analysis. These 50 signature genes are highly distinctive for these two subtypes. In the analysis, we pretended that the subtype information of each patient was unknown and evaluated the clustering accuracy of various clustering methods by comparing the estimated groups with the true subtypes. In all methods, we fixed $K = 2$. Moreover, we set the tuning parameters $\lambda_1 = 0.125$, $\lambda_2 = 0.3$, and $\lambda_3 = 0.12$ in our SCAN algorithm. For a fair comparison, we also used the same λ_1, λ_2 in Zhou et al. (2009), and the same λ_2, λ_3 in the joint graphical lasso of the two-stage method.

Table 3 reported the clustering errors of all methods as well as the number of informative variables in the corresponding estimated means and precision matrices. The standard *K*-means clustering has the largest clustering error due to its ignorance of the network structure in the precision matrices.

Therefore, the consequent joint graphical lasso method of the network reconstruction is less reliable. The method in [Zhou et al. \(2009\)](#) improves K -means clustering but is less accurate than ours. This is because their method estimates each precision matrix individually without borrowing information from each other. In this gene network example, both graphical models share many edges due to the commonality in the GBM diseases. Our method is able to achieve the best clustering performance due to the procedure of simultaneous clustering and heterogeneous graphical model estimation.

Table 3: The clustering errors and the number of selected features in cluster mean and precision matrix of various methods in the Glioblastoma Cancer Data.

Methods	Clustering Error	$\sum_k \ \hat{\boldsymbol{\mu}}^{(k)}\ _0$	$\sum_k \ \hat{\boldsymbol{\Omega}}^{(k)}\ _0$
K -means	0.203	100	NA
Zhou et al. (2009)	0.191	91	596
K -means + JGL	0.203	100	520
SCAN	0.149	70	564

To evaluate the ability of reconstructing gene regulatory network of each subtype, we report the two gene networks estimated from our SCAN method in Figure 1. The black lines are the links shared in both subtypes, and the thick red and thick green lines are uniquely presented in each subtype. Clearly, most edges are black lines, which indicates the common structure of both subtypes. There are two red links that only present in the Classical subtype: PTPRC \leftrightarrow DMWD, FBXO3 \leftrightarrow RAD21, and there are six green links that only presented in the Mesenchymal subtype: ICK \leftrightarrow LPR6, M6PRBP1 \leftrightarrow SIRT5, SEC61A2 \leftrightarrow M6PRBP1, SEMA6A \leftrightarrow SCAMP4, SEMA6A \leftrightarrow REXANK, SCAMP4 \leftrightarrow ROGDI. These findings agree with the existing results in [Verhaak et al. \(2010\)](#). It has been shown that the PTPRC is a well-described microglia marker and is highly exposed in the set of murine astrocytic samples which are strongly associated with the Classical group. In addition, as illustrated in Figure 2A in [Verhaak et al. \(2010\)](#), FBXO3 is informative in the Classical group and is less significant in the Mesenchymal group. Furthermore, as reported in Table S3 in [Verhaak et al. \(2010\)](#), SEC61A2 is a gene with the function of protein binding and SEMA6A is a gene on development process, and both of these two gene sets appear frequently in the Mesenchymal samples. It would also be of interest to further investigate the unique gene links that were not discovered in existing literatures for better understanding of the GBM diseases.

6 Discussion

In this paper, we propose a new SCAN method for simultaneous clustering and estimation of heterogeneous graphical models with common structures. In theory, we show that the estimation error bound of our SCAN algorithm consists of statistical error and optimization error, which explicitly addresses the trade-off between statistical accuracy and computational complexity. In our experiments, the tuning parameters can be chosen via an efficient BIC-type criterion. For future work, it is of interest to investigate the model selection consistency of these tuning parameters and study the distributed implementation of ECM algorithm based on the work in (Wolfe et al., 2008).

References

- BALAKRISHNAN, S., WAINWRIGHT, M. J. and YU, B. (2016). Statistical guarantees for the em algorithm: From population to sample-based analysis. *Annals of Statistics* To appear.
- BALCAN, M., LIANG, Y. and GUPTA, P. (2014). Robust hierarchical clustering. *Journal of Machine Learning Research* **15** 4011–4051.
- BICKEL, P. and LEVINA, E. (2008). Covariance regularization by thresholding. *Annals of Statistics* **36** 2577–2604.
- BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* **3** 1–122.
- CAI, T. T., LI, H., LIU, W. and XIE, J. (2016a). Joint estimation of multiple high-dimensional precision matrices. *Statistica Sinica* **26** 445–464.
- CAI, T. T., LIU, W. and ZHOU, H. H. (2016b). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Ann. Statist.* **44** 455–488.
- CHEN, Y., PAVLOV, D. and CANNY, J. (2009). Large-scale behavioral targeting. In *ACM SIGKDD*.
- DANAHER, P., WANG, P. and WITTEN, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 373–397.
- FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B* **70** 849–911.

- FRIEDMAN, J., HASTIE, H. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.
- GAO, C., ZHU, Y., SHEN, X. and PAN, W. (2016). Estimation of multiple networks in gaussian mixture models. *Electronic Journal of Statistics* **10** 1133–1154.
- GUO, J., LEVINA, E., MICHAELIDIS, G. and ZHU, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98** 1–15.
- HORN, R. A. and JOHNSON, C. R. (1988). *Matrix Analysis*. New York: Cambridge Univ. Press.
- JEZIORSKI, P. and SEGAL, I. (2015). What makes them click: Empirical analysis of consumer demand for search advertising. *American Economic Journal* **7** 24–53.
- LAURITZEN, S. (1996). *Graphical Models*. Oxford Science Publications.
- LEE, W. and LIU, Y. (2015). Joint estimation of multiple precision matrices with common structures. *Journal of Machine Learning Research* **16** 1035–1062.
- MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 281–297.
- MCLACHLAN, G. and KRISHNAN, T. (2007). *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics.
- MENG, X.-L. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika* **80** 267–278.
- NARASIMHAN, M., JOJIC, N. and BILMES, J. (2006). Q-clustering. *Advances in Neural Information Processing Systems* .
- NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of m -estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557.
- PAN, W. and SHEN, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research* **8** 1145–1164.
- PETERSON, C., STINGO, F. C. and VANNUCCI, M. (2015). Bayesian inference of multiple gaussian graphical models. *Journal of the American Statistical Association* **110** 159–174.

- QIU, H., HAN, F., LIU, H. and CAFFO, B. (2016). Joint estimation of multiple graphical models from high dimensional time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78** 487–504.
- ROTHMAN, A. J. and FORZANI, L. (2014). On the existence of the weighted bridge penalized gaussian likelihood precision matrix estimator. *Electronic Journal of Statistics* **8** 2693–2700.
- SAEGUSA, T. and SHOJAIE, A. (2016). Joint estimation of precision matrices in heterogeneous populations. *Electronic Journal of Statistics* To appear.
- SHEN, X., PAN, W. and ZHU, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association* **107** 223–232.
- SHOJAIE, A. and MICHAILIDIS, G. (2010a). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika* **97** 519–538.
- SHOJAIE, A. and MICHAILIDIS, G. (2010b). Penalized principal component regression on graphs for analysis of subnetworks. *Advances in Neural Information Processing Systems* 2155–2163.
- SUN, W., WANG, J. and FANG, Y. (2012). Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electron. J. Statist.* **6** 148–167.
- TCGA (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455** 1061–1068.
- VERHAAK, R. G., HOADLEY, K. A., PURDOM, E., WANG, V., QI, Y., WILKERSON, M. D., MILLER, C. R., DING, L., GOLUB, T., MESIROV, J. P., ALEXE, G., LAWRENCE, M., OKELLY, M., TAMAYO, P., WEIR, B. A., GABRIEL, S., WINCKLER, W., GUPTA, S., JAKKULA, L., FEILER, H. S., HODGSON, J. G., JAMES, C. D., SARKARIA, J. N., BRENNAN, C., KAHN, A., SPELLMAN, P. T., WILSON, R. K., SPEED, T. P., GRAY, J. W., MEYERSON, M., GETZ, G., PEROU, C. M., HAYES, D. N. and TCGA (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in *pdgfra*, *idh1*, *egfr*, and *nf1*. *Cancer Cell* **17** 98–110.
- VERSHYNIN, R. (2012). *Compressed sensing*, chap. Introduction to the non-asymptotic analysis of random matrices. Cambridge Univ. Press, 210–268.
- WAINWRIGHT, M. J. (2014). Structured regularizers for high-dimensional problems: Statistical and computational issues. *Annual Review of Statistics and Its Application* **1** 233–253.

- WANG, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika* **97** 893–904.
- WANG, J. (2015). Joint estimation of sparse multivariate regression and conditional graphical models. *Statistica Sinica* **25** 831–851.
- WANG, P., SUN, W., YIN, D., YANG, J. and CHANG, Y. (2015a). Robust tree-based causal inference for complex ad effectiveness analysis. In *Proceedings of 8th ACM Conference on Web Search and Data Mining*.
- WANG, Z., GU, Q., NING, Y. and LIU, H. (2015b). High dimensional em algorithm: Statistical optimization and asymptotic normality. *Advances in Neural Information Processing Systems* 2512–2520.
- WOLFE, J., HAGHIGHI, A. and KLEIN, D. (2008). Fully distributed em for very large datasets. *The International Conference on Machine Learning* 1184–1191.
- YAN, J., LIU, N., WANG, G., ZHANG, W., JIANG, Y. and CHEN, Z. (2009). How much can behavioral targeting help online advertising? In *International ACM WWW Conference*.
- YI, X. and CARAMANIS, C. (2015). Regularized em algorithms: A unified framework and statistical guarantees. *Advances in Neural Information Processing Systems* 1567–1575.
- YUAN, M. and LIN, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika* **94** 19–35.
- ZHANG, Y., CHEN, X., ZHOU, D. and JORDAN, M. I. (2014). Spectral methods meet em: A provably optimal algorithm for crowdsourcing. *Advances in Neural Information Processing Systems* 1260–1268.
- ZHOU, H., PAN, W. and SHEN, X. (2009). Penalized model-based clustering with unconstrained covariance matrices. *Electron. J. Statist.* **3** 1473–1496.
- ZHU, Y., SHEN, X. and PAN, W. (2014). Structural pursuit over multiple undirected graphs. *Journal of the American Statistical Association* **109** 1683–1696.